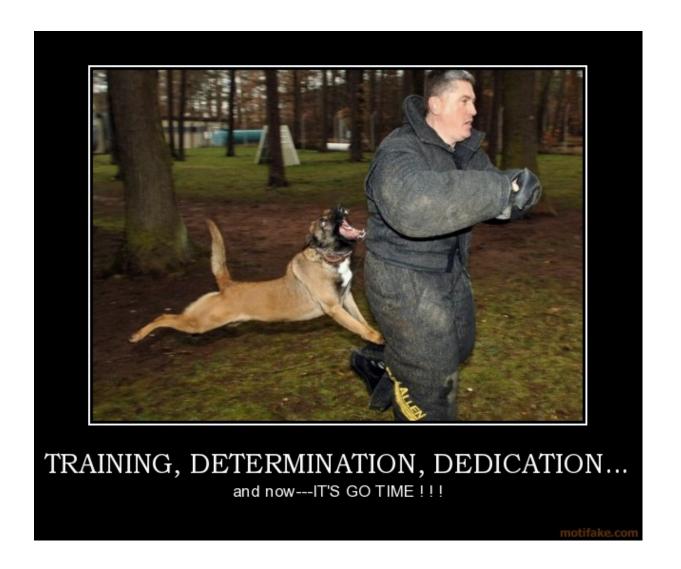
Metagenomic Analysis

An overview of the sequencing tools



The jargon

- Raw reads
- Unassembled vs assembled
- Read based
- K-mer
- Reference

The objective

- 16S rRNA amplicon
 - OTU table (counts per "species")
 - Classification ("species" taxonomy)
 - Metadata (sample specific information)

The objective

- Metagenomic
 - Gene/genome count table (counts per "species")
 - Classification (gene taxonomy/function)
 - Metadata (sample specific information)

16S rRNA amplicon sequencing

Pros:

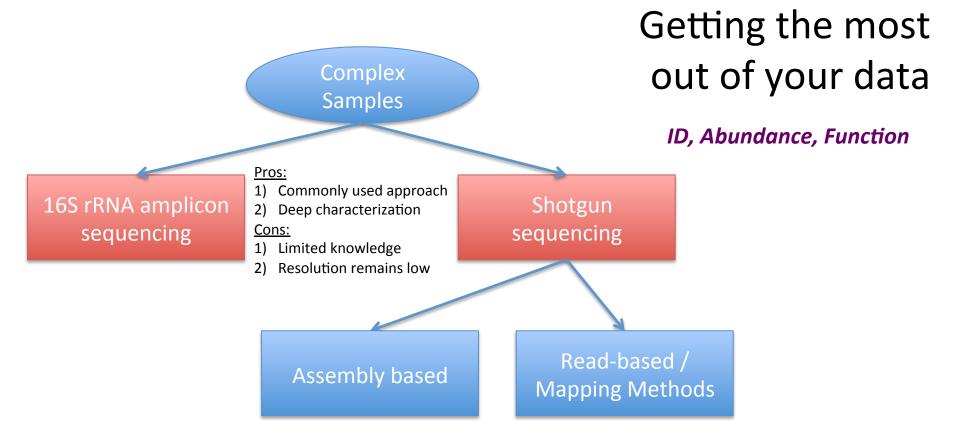
- 1) Commonly used approach
- 2) Deep characterization

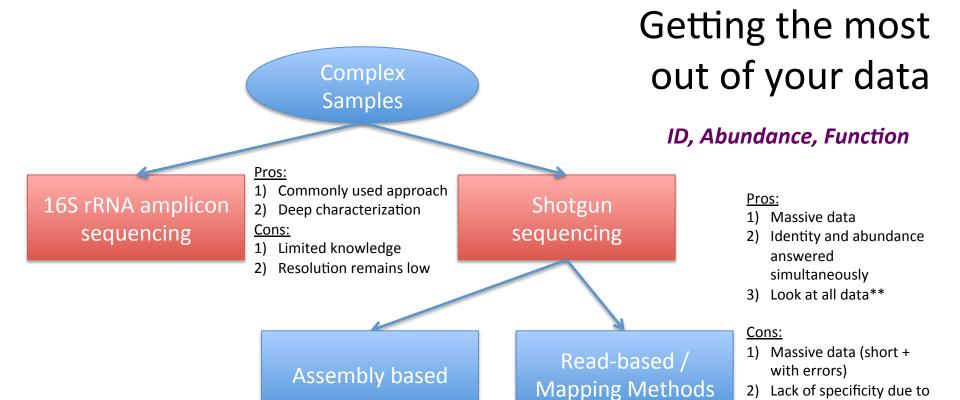
Cons:

- 1) Limited knowledge
- 2) Resolution remains low

Getting the most out of your data

ID, Abundance, Function

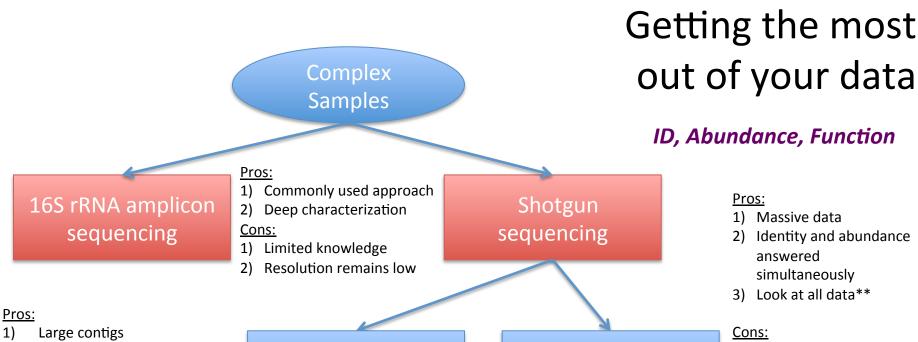




FPs from genomic

3) Difficult to detect novel genomes – must infer

redundancy



- 2) **Positional Information**
- 3) Most direct method to identify novel orgs/genes

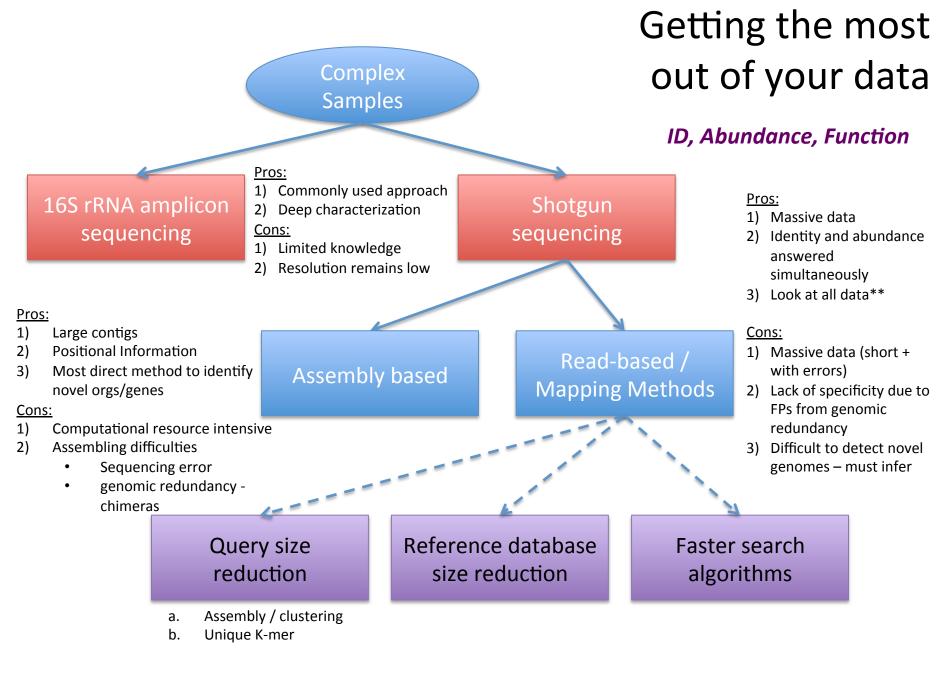
Cons:

- Computational resource intensive
- Assembling difficulties
 - Sequencing error
 - genomic redundancy chimeras

Assembly based

Read-based / Mapping Methods

- 1) Massive data (short + with errors)
- 2) Lack of specificity due to FPs from genomic redundancy
- 3) Difficult to detect novel genomes - must infer



Getting the most out of your data Complex Samples ID, Abundance, Function Pros: 1) Commonly used approach Pros: 16S rRNA amplicon Shotgun 2) Deep characterization 1) Massive data Cons: sequencing sequencing 2) Identity and abundance 1) Limited knowledge answered 2) Resolution remains low simultaneously 3) Look at all data** Pros: Large contigs Cons: **Positional Information** 2) 1) Massive data (short + Read-based / 3) Most direct method to identify with errors) Assembly based Mapping Methods novel orgs/genes 2) Lack of specificity due to FPs from genomic Cons: Computational resource intensive redundancy Assembling difficulties 3) Difficult to detect novel Sequencing error genomes - must infer genomic redundancy chimeras Reference database Query size Faster search reduction size reduction algorithms Assembly / clustering a. Selection of marker a. b. Unique K-mer genes Identification of b. signatures (Kmers)

Getting the most out of your data

ID, Abundance, Function

16S rRNA amplicon sequencing

Pros:

- 1) Commonly used approach
- 2) Deep characterization
- Cons:
- 1) Limited knowledge
- 2) Resolution remains low

Shotgun sequencing

Pros:

- 1) Massive data
- Identity and abundance answered simultaneously
- 3) Look at all data**

Pros:

- Large contigs
- 2) Positional Information
- 3) Most direct method to identify novel orgs/genes

Cons:

- 1) Computational resource intensive
- 2) Assembling difficulties
 - Sequencing error
 - genomic redundancy chimeras

Assembly based

Read-based / Mapping Methods

Cons:

- Massive data (short + with errors)
- 2) Lack of specificity due to FPs from genomic redundancy
- 3) Difficult to detect novel genomes must infer

Query size reduction

- a. Assembly / clustering
- b. Unique K-mer

Reference database size reduction

- a. Selection of marker genes
- b. Identification of signatures (Kmers)

Faster search algorithms

- a. Exact match
- b. K-mer based search
- c. Improved algorithm
 - a. Clustering
 - b. Pattern matching

Getting the most out of your data

ID, Abundance, Function

16S rRNA amplicon sequencing

Pros:

- 1) Commonly used approach
- 2) Deep characterization

Cons:

- 1) Limited knowledge
- 2) Resolution remains low

Shotgun sequencing

Pros:

- 1) Massive data
- Identity and abundance answered simultaneously
- S) Look at all data**

Pros:

- Large contigs
- 2) Position
- a) Most di Tutorial One: Quality trim youi novel oi unassembled reads

Cons:

- 1) Computational resource intensive
- 2) Assembling difficulties
 - Sequencing error
 - genomic redundancy chimeras

Query size reduction

- a. Assembly / clustering
- b. Unique K-mer

Read-based / Mapping Methods

Cons:

- Massive data (short + with errors)
- Lack of specificity due to FPs from genomic redundancy
- Difficult to detect novel genomes must infer

Reference database size reduction

- a. Selection of marker genes
- b. Identification of signatures (Kmers)

Faster search algorithms

- a. Exact match
- b. K-mer based search
- c. Improved algorithm
 - a. Clustering
 - b. Pattern matching

Getting the most out of your data

ID, Abundance, Function

16S rRNA amplicon sequencing

Pros:

1) Commonly used approach

Assembly based

- 2) Deep characterization
- Cons: 1) Limited knowledge
- 2) Resolution remains low

Shotgun sequencing

Pros:

- 1) Massive data
- 2) Identity and abundance answered simultaneously
- 3) Look at all data**

Pros:

- Large contigs
- **Positional Information** 2)
- 3) Most direct method to identify novel orgs/genes

Cons:

- Computational resource intensive
- Assembling difficulties
 - Sequencing error
 - genomic recondance chimeras

Read-based / Mapping Methods

and assemble

Cons:

- Massive data (short + with errors)
- 2) Lack of specificity due to FPs from genomic redundancy Difficult to detect novel genomes – must infer

Query size reduction

- Assembly / clustering a.
- Unique K-m r

Reference database size reduction

- Selection of marker genes
- Identification of b. signatures (Kmers)

Faster search algorithms

- Exact match
- b. K-mer based search
- Improved algorithm
 - Clustering a.
 - b. Pattern matching

Getting the most out of your data

ID, Abundance, Function

16S rRNA amplicon sequencing

Pros:

- 1) Commonly used approach
- 2) Deep characterization

Cons:

- 1) Limited knowledge
- 2) Resolution remains ov

Shotgun sequencing

Pros:

- 1) Massive data
- Identity and abundance answered simultaneously
- 3) Look at all data**

Pros:

- Large contigs
- 2) Positional Information
- Most direct method to identify novel orgs/genes

Cons:

- 1) Computational resource intervive
- 2) Assembling difficulties
 - Sequencing error
 - genomic redundancy chimeras

Assembly based

Read-based / Mapping Methods

Tutorial Three: assess your assembly

Cons:

- Massive data (short + with errors)
- 2) Lack of specificity due to FPs from genomic redundancy Difficult to detect novel genomes must infer

Query size reduction

- a. Assembly / clustering
- b. Unique K-mer

Reference database size reduction

- a. Selection of marker genes
- b. Identification of signatures (Kmers)

Faster search algorithms

- a. Exact match
- b. K-mer based search
- c. Improved algorithm
 - a. Clustering
 - b. Pattern matching

Getting the most out of your data

ID, Abundance, Function

16S rRNA amplicon sequencing

Pros:

- 1) Commonly used approach
- 2) Deep characterization
- Cons:
- 1) Limited knowledge
- 2) Resolution remains low

Shotgun sequencing

Pros:

- 1) Massive data
- 2) Identity and abundance answered simultaneously
- 3) Look at all data**

Pros:

- Large contigs
- **Positional Information** 2)
- 3) Most direct method to identify novel orgs/genes

Cons:

- Computational resource intensive
- Assembling difficulties
 - Sequencing error
 - genomic redundancy chimeras

Assembly based

Read-based / Mapping Methods

estimate abundances

Cons:

- 1) Massive data (short + with errors)
- 2) Lack of specificity due to FPs from genomic redundancy.
 - Difficult tetect novel genomes - hust infer

Query size reduction

- Assembly / clustering a.
- b. Unique K-mer

Reference database size reduction

- Selection of marker a. genes
- Identification of b. signatures (Kmers)

Faster search algorithms

Exact match a.

a.

- K-mer based so arch
 - Liproyed and rithm Clustering
 - b. Pattern matching

Getting the most out of your data

ID, Abundance, Function

16S rRNA amplicon sequencing

Pros:

- 1) Commonly used approach
- 2) Deep characterization

Cons:

- 1) Limited knowledge
- 2) Resolution remains low

Shotgun sequencing

Read-based /

Pros:

- 1) Massive data
- Identity and abundance answered simultaneously
- 3) Look at all data**

Pros:

- Large contigs
- 2) Positional Information
- 3) Most direct method to identify novel orgs/genes

Cons:

- 1) Computational resource intensive
- 2) Assembling difficulties
 - Sequencing error
 - genomic redundancy chimeras

Assembly based

Tutorial Five: Identifying signatures (binning by abundances)

Cons:

- Massive data (short + with errors)
- 2) Lack of specificity due to FPs from genomic redundancy
- 3) Difficult to detect novel genomes must infer

Query size reduction

- a. Assembly / clustering
- b. Unique K-mer

eference database size reduction

- a. Selection of marker genes
- b. Identification c signatures (Kniers)

Faster search algorithms

- a. Exact match
- b. K-mer based search
- c. Improved algorithm
 - a. Clustering
 - b. Pattern matching

Getting the most out of your data

ID, Abundance, Function

16S rRNA amplicon sequencing

Pros:

- 1) Commonly used approach
- 2) Deep characterization

Cons:

- 1) Limited knowledge
- 2) Resolution remains low

Shotgun sequencing

Pros:

- 1) Massive data
- Identity and abundance answered simultaneously
- 3) Look at all data**

Pros:

- Large contigs
- 2) Positional Information
- 3) Most direct method to identify novel orgs/genes

Cons:

- 1) Computational resource intensive
- 2) Assembling difficulties
 - Sequencing error
 - genomic redundancy chimeras

Assembly based

Read-based / Mapping Methods

Cons:

- Massive data (short + with errors)
- 2) Lack of specificity due to FPs from genomic redundancy
- 3) Difficult to detect novel genomes must infer

Query size reduction

- a. Assembly / clustering
- b. Unique K-mer

Reference database size reduction

- a. Selection of marker genes
- b. Identification of signatures (Kmers)

Faster search algorithms

- a. Exact match
- b. K-mer based search
- c. Improved algorithm
 - a. Clustering
 - b. Pattern matching

So let's do some learning...

